The Stable Signature: Rooting Watermarks in Latent Diffusion Models

Year: 2023 | Citations: 279 | Authors: Pierre Fernandez, Guillaume Couairon, Herv'e J'egou, Matthijs Douze, T. Furon

Abstract

Generative image modeling enables a wide range of applications but raises ethical concerns about responsible deployment. We introduce an active content tracing method combining image watermarking and Latent Diffusion Models. The goal is for all generated images to conceal an invisible watermark allowing for future detection and/or identification. The method quickly fine-tunes the latent decoder of the image generator, conditioned on a binary signature. A pre-trained watermark extractor recovers the hidden signature from any generated image and a statistical test then determines whether it comes from the generative model. We evaluate the invisibility and robustness of the watermarks on a variety of generation tasks, showing that the Stable Signature is robust to image modifications. For instance, it detects the origin of an image generated from a text prompt, then cropped to keep 10% of the content, with 90+% accuracy at a false positive rate below 10–6.